IDIAP
Martigny - Valais - Suisse

IDIAP RESEARCH REPORT

ASYMMETRIC FILTER FOR TEXT RECOGNITION IN VIDEO

Datong Chen, Kim Shearer
IDIAP
Case Postale 592
Martigny Switzerland

IDIAP–RR 00-37

Nov. 2000

Institut Dalle Molle d'Intelligence Artificielle Perceptive • CP 592 • Martigny • Valais • Suisse

téléphone +41 – 27 – 721 77 11
télécopieur +41 – 27 – 721 77 12
adr.él. secretariat@idiap.ch
internet http://www.idiap.ch

Rapport de recherche de l'IDIAP 00-37

ASYMMETRIC FILTER FOR TEXT RECOGNITION IN VIDEO

Datong Chen, Kim Shearer
IDIAP
Case Postale 592
Martigny Switzerland

Nov. 2000

tripes are a common sub-structure of text characters, and the scale of the stripes does not vary significantly within a character. In this paper a new form of filter is derived from the Gabor filter which can efficiently estimate the scales of such stripes. The contrast of text in video can then be increased by enhancing the edges of those stripes found to have a suitable scale. The algorithm presented enhances the stripes in three selected scale ranges. Character recognition is then performed on the output of binarizing these enhanced images, and shows improvement over other methods.

1 Introduction

Video text detection and recognition is a useful technology for building text based video indexing systems. Video text including captions and embodied scene text, provides precise and explicit information about the names of people, organization, location, date, times and scores, etc., which are powerful resources for video indexing and annotation. The idea of combining the text-based searching technologies and text recognition technologies has become a popular solution of the video indexing in recent years [11][12][13]. As text-based searching and retrieval technologies have been well developed [1][2], there is high demand for a system to detect and recognize unconstrained text against any background in video.

In constrast with OCR text printed against a clean background, video text may have the background of many kinds of indoor or outdoor scenes. Therefore, some properties of the video text, such as scale and alignment, can not be obtained easily unless the text can be segmented from the video frames. In order to address this problem, much former work presents the methods based on region analysis and texture analysis.

Region-based methods, first, roughly segment the input video frames into regions. The text regions are then selected by checking the features, such as maximum and minimum of the region, the width/height ratio of the region, contrast between the segments and the background, spatial frequency of the region, alignment and movement [4][11][9][10]. The region-based text detection methods require that the gray-level of the text should be consistent and can not improve the segmentation result obtained at the first step of the algorithm.

Texture-based methods use the texture property of the text string to search for text in the video. Wu et.al. [5][6][7] proposed an algorithm using statistical properties of the pixel values through out the Gaussian scale space to segment the input image into text regions and background. The candidate text regions are refined under heuristic constraints, such as height similarity, spacing and linear alignment and fixed spatial frequency. Some texture features of the text string can also be found in single scale images. Zhong [8] uses spatial variance of the input image as a feature for locating the text. In [14], the texture feature is extracted using Haar wavelet transform. The texture-based methods are robust to small gray-level variation of the text but can not always locate the text accurately [15].

In this paper, we presents a method for detecting and segmenting the text using local substructures. We locate the sub-structures of the text using edge detection and estimate the scale of each sub-structures using a family of filters, which is presented in section 2. We then select three scale ranges and enhance the contrast of these sub-structures as character strokes in each scale range individually to improve the performance of text segmentation.

2 Asymmetric Filter

The family of two-dimensional Gabor filters $G_{\lambda,\theta,\varphi}(x,y)$, which was proposed by Daugman [16], are often used to obtain the spatial frequency of the local pattern in an image:

$$G_{\lambda,\theta,\varphi}(x,y) = e^{-\frac{\left(x'^2+\gamma^2 y'^2\right)}{2\sigma^2}} cos\left(2\pi\frac{x'}{\lambda}+\varphi\right)$$

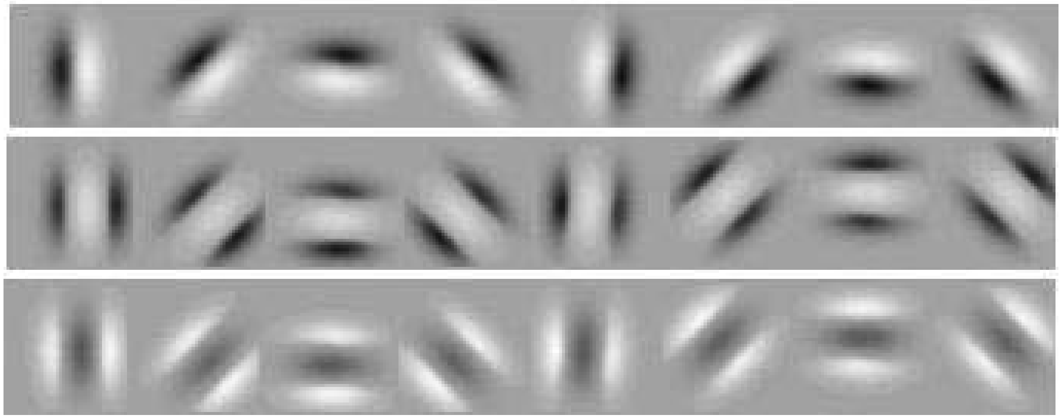

FIG. 1 – *Asymmetric filters*

$$\begin{aligned} x' &= xcos\theta + ysin\theta \\ y' &= -xsin\theta + ycos\theta \end{aligned}$$

Here the arguments x and y specify the position of the pixels over the image domain, the parameter θ specifies the orientation of the filter, the parameter γ determines the spatial aspect ratio, and $\frac{1}{\lambda}$ is called the spatial frequency. These filters provide the conjointly optimal resolution for the orientation, spatial frequency and the position of the local image structure.

However, text detection requires the precise position and scale of the text. We propose a family of asymmetric filters based on Gabor filters to obtain the precise scale information of the local image structures. Our family of filters is designed into two groups: edge-form filters and stripe-form filters. The edge-form filters $E_{\lambda,\theta}(x,y)$ are the Gabor filters with $\varphi = \pi/2$:

$$E_{\lambda,\theta}(x,y) = e^{-\frac{\left(x'^2+\gamma^2 y'^2\right)}{2\sigma^2}} cos\left(2\pi\frac{x'}{\lambda}+\frac{\pi}{2}\right)$$

$$\begin{aligned} x' &= xcos\theta + ysin\theta \\ y' &= -xsin\theta + ycos\theta \end{aligned}$$

The stripe-form filters $S_{\lambda,\theta}(x,y)$ are defined as a Gabor filter with a translation $\left(-\frac{\lambda}{4},0\right)$:

$$S_{\lambda,\theta}(x,y) = e^{-\frac{\left(x'^2+\gamma^2 y'^2\right)}{2\sigma^2}} cos\left(2\pi\frac{x'}{\lambda}\right)$$

$$\begin{aligned} x' &= xcos\theta + ysin\theta - \tfrac{\lambda}{4} \\ y' &= -xsin\theta + ycos\theta \end{aligned}$$

The edge-form and stripe-form filters keep most of the properties of the Gabor filters except the specified translation on the position and the phase offset. Figure 1 shows the pattern of the edge-form filters and stripe-form filters in 8 orientations with $\gamma = 0.92$.

A special property of these two groups of filters is very useful to find out the scale of the local image structure. Experiments show that if the pixel (x,y) is on the edge of stripe structure, the responses

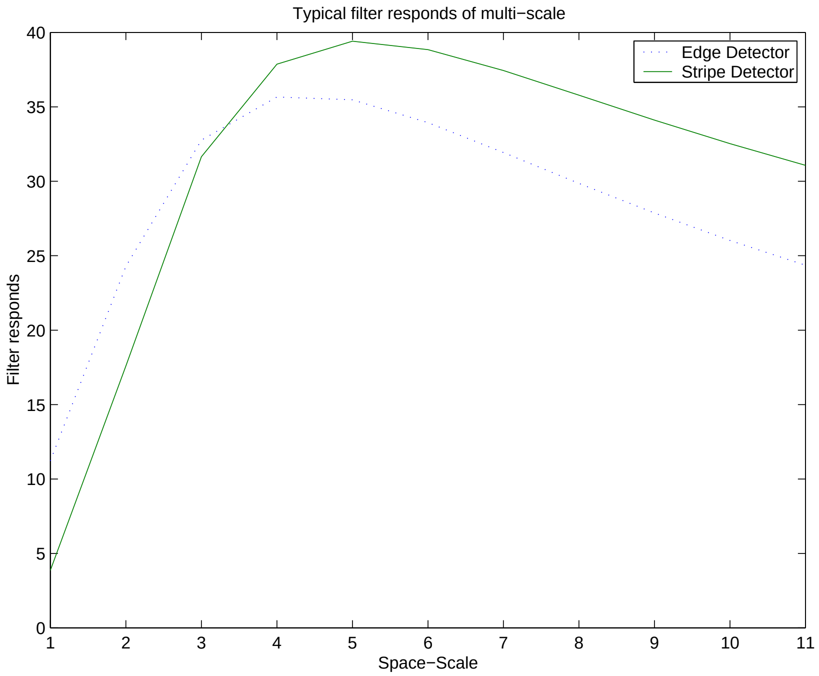

Fig. 2 – *scale adaptive property of asymmetric filters*

of the stripe-form filters are smaller than the responses of the edge-form filters when the scale of the filters are rather smaller than the scale of the stripe ($S_{\lambda,\theta}\,(x,y) < E_{\lambda,\theta}\,(x,y)$), but greater when the scale of the filters are rather larger than the scale of the stripe ($E_{\lambda,\theta}\,(x,y) < S_{\lambda,\theta}\,(x,y)$, see figure 2). The scale which the stripe-form filter response surpasses the edge-form filter response is called the surpass scale.

For any pixels on the edges, according to the surpass scale λ_s, we compute the filter responses of the two other scales $\frac{1}{2}\lambda_s$ and $\frac{3}{2}\lambda_s$. A neural network is trained with these five filter responses (the responses of the edge-form filter and the stripe-form filter at surpass scale are the same value) to find the scale of the local edge.

3 Algorithm

An off-line text-based video indexing system consists of text annotation and a text search engine. The search engine queries the text annotation with keywords input by the user and retrives the matching video periods.

3.1 Text Detection and Recognition

The text recognition algorithm can be summarized as follows:

1. The edges in the input video frame are extracted using the Canny algorithm.

2. In order to select the candidate pixels on the edges, we segment the frame in to $n \times n$ blocks, where n equals to the half size of the smallest scale of the substructures of the text (here n=2). Only one edge point is selected as the candidate pixel in each block. The pixle chosen is the one with the maximum energy.

3. The scales of the candidate pixels are estimated with the asymmetric filters and neutral network that we discussed in the last section. The pixels with the scales extremely small or large are filtered out because these values are out of our searching range. In this paper, the scale range is 3 to 50.

TAB. 1 – *Recognition results: the region analysis algorithm emploies MoCA algorithm without motion analysis. Text 1: superimposed text. Text 2: scene text*

algorithm	text	frames	recognition rate
MoCA	1	4000	36.1%
MoCA	2	4000	7.4%
Ours	1	4000	70.6%
Ours	2	4000	11.4%

4. With the filters in their scales and orientations of the candidate pixels, we can reconstruct the image pattern. The image pattern is reconstructed in three scale levels. The first level L_1 includes the scales between 3 and 9, the second level L_2consists of the scales from 7 to 30 and the last scale level L_3 is from 26 to 50. For each candidate pixel (x_i,y_i), with the orientation θ_i and filter scale λ_i, the three reconstructed image patterns are defined as:

$$I_r^k = \sum_{i=0}^{n} F^k(i),\ k = 1,2,3.$$

$$F^k(i) = \begin{cases} 0 & \lambda_i \notin L_k \\ S_{\lambda_i,\theta_i} & \lambda_i \in L_k \end{cases} .$$

5. The original image is then enhanced by addition of itself to the three reconstructed image patterns I_r^k individually,

$$I_e^k(x,y) = I_{org}(x,y) + I_r^k(x,y),\ k = 1,2,3.$$

We then binarize these three enhanced images using the Otsu's method and mask off the pixels with zero value in the reconstructed image patterns. These three binary image are used as the inputs of a commercial OCR software system to recognize the text characters.

4 Experiment and Result

Experiments are based on 3 video streams with a total of 4000 frames (512x512). The text in the videos is both superimposed text and scene text, with different alignments and movements. When applying our algorithm, we constrain that the minimum scale of the sub-structure of the text to 3 pixels and the maximum scale to 50 pixels. Text characters are assumed to appear in at least 8 consecutive frames. The examples of image patterns, enhanced images and the binary images in three scale ranges and binary result of MoCA are shown in figure 3. The final recognition results with the TypeRead OCR package can be found in table 1, which shows the comparision between results of our algorithm and the results with MoCA.

5 Conclusion

Stripes are a very common sub-structure of characters. Their scale does not vary much in one text character. Using our family of asymmetric filters, we can estimate the scale of a stripe efficiently with a binary search algorithm. It is therefore possible to enhance contrast at only those edges most likely to represent text in the scale interest of. The experimental results show that the approach presented in this paper can improve the recognition rate of the superimposed text significantly.

The main disadvantage of this method is that it can not filter out all background from the frames. Some local structure of the background may also contain similar stripe structure. They are regarded as part of text characters and let OCR system make the final decision.

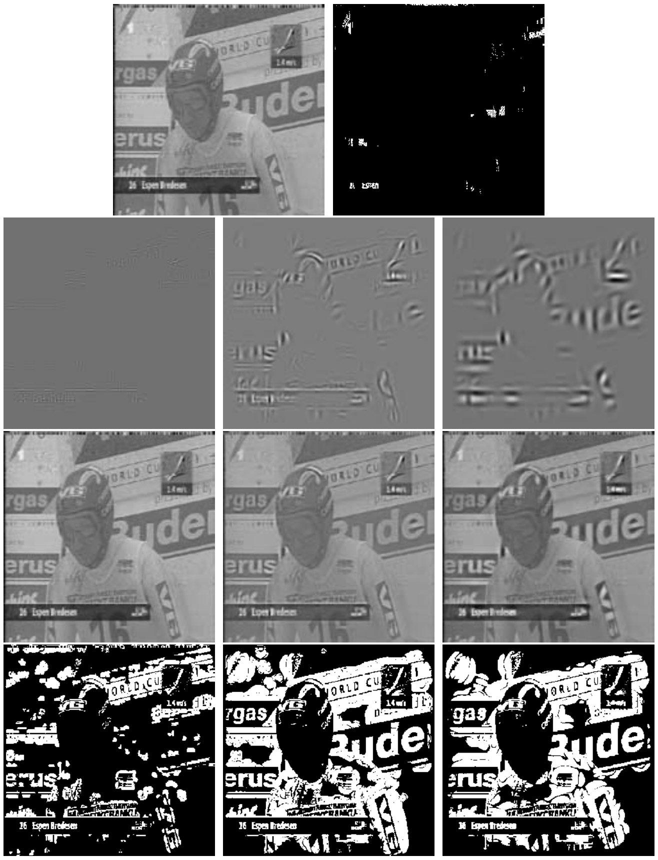

Fig. 3 – *row 1: original frame image (left), MoCA result (right); (row 2) reconstructed image patterns in 3 scale ranges; (row 3) enhanced images in 3 scale ranges; (row 4) binarization results of 3 enhanced images, with the Otsu's method. The images from left to right are* L_1,L_2,L_3

Références

[1] M. Bokser, "Omnidocument technologies", Proc. IEEE, 80(7):1066–1078, July 1992.

[2] S. V. Rice, F. R. Jenkins, and T. A. Nartker. "OCR accuracy: UNLV's fifth annual test", INFORM, 10(8), September 1996.

[3] L. O'Gorman and R. Kasturi, "Document Image Analysis", IEEE Computer Society Press, Los Alamitos, 1995.

[4] J. Ohya, A. Shio, and S. Aksmatsu, "Recognition characters in scene images. IEEE Trans. Pattern Analysis and Machine Intelligence", 16(2):214–220, 1994.

[5] V. Wu, R. Manmatha, and E. M. Riseman, "Finding text in images", In Proc. ACM Int. Conf. Digital Libraries, 1997.

[6] V. Wu, R. Manmatha, and E. M. Riseman, "Textfinder: An automatic system to detect and recognize text in images", IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11):1224–1229, 1999.

[7] V. Wu and R. Manmatha, "Document image clean-up and binarization", In Proc. SPIE Symposium on Electronic Imaging, 1998.

[8] Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color images", Pattern Recognition, 28(10):1523–1536, 1995.

[9] R. Lienhart and W. Effelsberg, "Automatic text segmentation and text recognition for video indexing", Technical Report TR-98-009, University of Mannheim, Mannheim, 1998.

[10] R. Lienhart, "Automatic text recognition in digital videos", In Proc. SPIE, Image and Video Processing IV, January 1996.

[11] R. Lienhart, "Indexing and retrieval of digital video sequences based on automatic text recognition", In Proc. 4th ACM International Multimedia Conference, Boston, November 1996.

[12] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video ocr for digital news archives", In IEEE Workshop on Content Based Access of Image and Video Databases, Bombay, January 1998.

[13] T. Sato, T. Kanade, E. K. Hughes, M. A. Smith, and S. Satoh, "Video OCR: indexing digital news libraries by recognition of superimposed caption", In ACM Multimedia System Special Issue on Video Libraries, Feb. 1998.

[14] H. Li and D. Doermann, "Text enhancement in digital video using multiple frame integration", ACM Multimedia 1999.

[15] A. K. Jain and B. Yu, "Automatic text localization in images and video frames", Pattern Recognition, 31(12):2055–2076, 1998.

[16] G. Daugman, "Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters", Journal of the Optical Society of America A, 1985, vol.2, pp. 1160-1169.